

Reconstruction of gene regulatory networks from steady state data

Arne B. Gjuvsland^{*,1} and Erik Plahte²

¹*Centre for Integrative Genetics, Dept. of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway*

²*Centre for Integrative Genetics, Dept. of Mathematical Sciences and Technology, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway*

Abstract

Genes are connected in regulatory networks, often modelled by ordinary differential equations. Changes in expression of a gene propagate to other genes along paths in the network. At a stable state, the system's Jacobian matrix confers information about network connectivity. To disclose the functional properties of genes, knowledge of network connections is essential. We present a new method to reconstruct the Jacobian matrix of models for gene regulatory systems from equilibrium protein concentrations. In a recent paper we defined propagation and feedback functions describing how genetic variation at one gene propagates to the other genes in the network and possibly also back to itself. Here we show how propagation and feedback functions provide relations between equilibrium protein levels which are in principle observable, and Jacobi elements which are not directly observable. We establish exact formulae expressing the Jacobian in terms of derivatives of propagation and feedback functions. Approximating these derivatives from perturbed and unperturbed protein levels, we derive formulae for estimating the Jacobian. We apply the method to models of the *Drosophila* segment polarity network and randomly generated gene networks. Genes could be perturbed in two ways: by modifying mRNA degradation rates, or by allele knockout in diploid models. Comparison with the true Jacobians shows that for noiseless data we obtain hit rates close to 100% in the former case and in the range 80-90% in the latter. Our method adds to the network interference toolbox and provides a sign estimate of the Jacobian from steady state data, and a value estimate of the Jacobian if protein degradation rates are known. Also the approach identifies some predicted connections as much more reliable than others, and could point to further experiments for resolving uncertainties in the less dependable Jacobian elements.

Keywords: Gene regulatory networks; Reverse engineering; Jacobian; Feedback

^{*}Email:arne.gjuvsland@nmbu.no

Background

Living organisms contain large numbers of complex networks in which genes, mRNA molecules, protein, metabolites etc. interact to maintain essential functions and to react to a wide range of external impacts and conditions. Genes interact when the protein output of a gene enters into the cell's biochemical system, and through interactions with other chemical species, frequently by long and intricate pathways, influence the expression of other genes by enhancing or inhibiting transcription or modifying translation. Thus a gene is not an independent object whose functions are only dependent on its internal structure and properties and the external conditions, but is part of a network, reacts to input from other genes in the network and in turn affect other components of the network (Emmert-Streib and Dehmer, 2011). Feedback loops and feedforward motifs are important building blocks in gene networks (Alon, 2007).

Systems of ordinary differential equations are frequently used to model the dynamics of such networks. In a dynamic system with a stable point, all solutions within its basin of attraction usually decay exponentially towards the stable point, which is then called hyperbolic. Close to it, dynamic trajectories can be computed approximately once the Jacobian matrix J is known. The Jacobian matrix also confers information about the network structure of the system: all the system's actions and interactions that are operative in the neighbourhood of the stable point, can be read out from its elements.

For a system for which no validated model exists, a basic question is whether it is possible to obtain information on the network topology and connections, which are not directly observable, from the concentrations of mRNA and proteins, which are directly observable. In the literature one finds a large number of papers describing different *reverse engineering* methods, see e.g. reviews by Brazhnik (2005); Camacho *et al.* (2007); Cho *et al.* (2007); Emmert-Streib *et al.* (2012); Goutsias and Lee (2007); Ross (2008); Stark *et al.* (2003a,b); Tirosh and Barkai (2011); Yip *et al.* (2010); Chai *et al.* (2014). There are essentially two main classes of methods: using time series data, or equilibrium concentrations. The present paper belongs to the latter class.

By repeatedly perturbing the system to induce a shift of equilibrium values of the state variables, one may hopefully be able to infer the Jacobian matrix elements. The observation that expression of gene B is affected by a perturbation of gene A does not in itself say much about J . From this fact alone one cannot tell whether the effect is direct or mediated by one or several other genes. This enigma might be solved by performing more perturbations, but by an *ad hoc* procedure one soon gets lost. A systematic approach is necessary.

The present paper is a continuation of our recent analysis of propagation of genetic variation in diploid networks (Plahte *et al.*, 2013). There we developed a formalism for describing how a change of genotype of one gene in a gene regulatory network propagates to other, downstream genes and possibly also back to the gene itself, and how this propagation effect is related to the structure of the network.

In the present paper we consider the opposite situation in which so far no network model for some gene regulatory system exists. How, and to what extent, can allele knockouts or other perturbations yield information by which a model of the network can be constructed?

In the following the term “gene” should be considered as a *functional module*, “an entity of known/unknown genes, proteins or metabolites, grouped together and internally connected by complex physico-chemical interactions” (Yalamanchili *et al.*, 2006). A module may take inputs from many other modules, including

itself, but we assume each module is delimited such that it only produces a single output. The process of obtaining information about the local interactions, i.e. the direct effect of a perturbation of one module on another, from the global effects that result from the web of network connections between the modules, has been called *Modular Response Analysis (MRA)* (Sontag, 2008), and was developed by Kholodenko *et al.* (2002) (see also Andrec *et al.* (2005); Cho *et al.* (2005); Yalamanchili *et al.* (2006)). Albeit similar, our approach differs from the original MRA in several respects. In the original MRA, the authors were only able to determine the rows of the Jacobian up to a scalar multiple. This shortcoming is related to the fact that the equilibrium conditions are unchanged if each rate function is multiplied with a nonzero constant, while the Jacobian is not. However, if the MRA is supplemented by non-steady state data, the full Jacobian can be estimated (Sontag *et al.*, 2004). Using only steady state data, we are able to determine the correct sign of the elements of the Jacobian, and if the protein degradation rates are known, their numeric values as well.

We consider two particular approaches to modify or perturb a gene and by that modifying the functioning of the network. The first approach involves perturbing the mRNA degradation rates, for instance through RNA interference methods. The second approach is to knock out one of the two alleles of a diploid gene. We illustrate both approaches by *in silico* experiments. We show that if the protein degradation rates can be measured or estimated, both approaches can be used to infer the Jacobian from unperturbed and perturbed protein concentration data from which the Jacobian of the system can be inferred.

Mathematical analyses and results

Model framework of gene regulatory networks

We consider a set of genes believed to be part of a network \mathcal{N} of n nodes or loci X_i , $i \in N$, where $N = \{1, \dots, n\}$ and $n \geq 2$. A real gene regulatory network can be modelled by a dynamic model designed according to the following lines. A non-negative variable z_i describes the concentration or amount of the output of X_i , its time rate of change being given by

$$\dot{z}_i = r_i(z, a_i) - \gamma_i z_i, \quad i \in N, \quad (1)$$

where $z = [z_1, \dots, z_n]$. The differentiable rate function $r_i(z, a_i)$ represents the production rate or dose-response function of X_i , and γ_i is its constant relative degradation rate. The quantity $a = \{a_i\}$, $i \in N$, represents a set of parameters defining the system's genotype, the subset a_i defining the genotype of X_i . This model framework is commonly used to model gene regulatory networks. In fact, it has been around for decades (de Jong, 2002). Frequently, the dose-response functions are modelled by means of sigmoidal functions, for example the well-known Hill function. We assume that Eqs. (1) have a single, hyperbolic equilibrium point x . Our goal is to estimate the Jacobian J of Eqs. (1) in x in terms of experimentally observable quantities.

Notation: Vector and matrix components are indicated by subscripts as usual. Superscripts are used extensively as indices and except in a few obvious cases, never indicate a power. For vectors and matrices, superscripts in parentheses indicate that the enclosed component has been excluded. A superscript (kj) to a matrix, for example $A^{(kj)}$, indicates that row number k and column number j in A have been deleted. Similarly, $a^{(k)}$ is the set (or vector) a with a_k deleted, $a^{(k)} = a \setminus \{a_k\}$. Superscripts in brackets indicate the value when a node has been perturbed. For example, $x_j^{[k]}$ is the equilibrium value of X_j when node X_k has been

perturbed. We also use a set notation for subscripts. For example, if L is a subset of N , then x_L is the vector with components x_l , $l \in L$. If A is a matrix, A_i denotes row number i in A . The equilibrium condition of X_j is denoted by E_j . The Jacobian of Eq. (1) is J , $D = \det(J)$ and $D^{(ij)} = \det(J^{(ij)})$. If M is a square matrix, $\text{diag}(M)$ is the diagonal matrix with the same main diagonal as M . The superscript T to a matrix denotes its transpose.

Eq. (1) could be seen as a simplification of a more realistic regulatory system model comprising mRNA, proteins and metabolites. Often gene outputs do not act directly as transcription factors regulating the expression rates of the genes. Rather, there are frequently long and complicated pathways that propagate and modify the regulatory processes. A philosophy behind Eq. (1) is that all these complicated reactions can be condensed into the response functions r_i (Brazhnik *et al.*, 2002). The segment polarity network model of von Dassow *et al.* (2000) is an example of a model framework in which the concentrations of mRNA and protein for each gene in the network are modelled independently. This model framework, which has been used by a number of authors (see e.g. Lewis (2003); Ichinose *et al.* (2008); Polynikis *et al.* (2009)), is

$$\begin{aligned}\dot{P}_i &= \rho_i m_i - \gamma_i P_i, \\ \dot{m}_i &= R_i(P) - \mu_i m_i,\end{aligned}\tag{2}$$

$i \in N$. Here P_i and m_i are the concentrations of protein and mRNA of gene number i , respectively, R_i is the production rate (dose-response function) of mRNA, dependent on the concentration of the input proteins, ρ_i is the mRNA-protein conversion rate, and γ_i and μ_i are positive relative degradation rates. The gene products might act directly as transcription factors, or the function $R_i(P)$ might implicitly contain chains of reactions from the gene products to the real transcription factors so that R_i is the combined effect of these chains and the transcription.

Of course the mRNA-protein conversion rate might not be constant, but some nonlinear function of m . However, our analysis is local around the stable point x , and the second of Eqs. (2) would then represent a locally valid linear approximation to the nonlinear transcription model.

As mRNA molecules are in general less stable than the corresponding protein molecules, we can safely assume that for all i , $\gamma_i \ll \mu_i$. With a number of reasonable assumptions we can make the quasi-stationarity hypothesis $\dot{m}_i \approx 0$, $m_i = R_i(x)/\mu_i$. This can be justified in a rigorous way by means of singular perturbation theory (see Appendix A), and leads to *the reduced model*

$$\dot{z}_i = \frac{\rho_i}{\mu_i} R_i(z) - \gamma_i z_i,\tag{3}$$

where $z = P$. This equation is of the general form Eq. (1) on which all our derivations are based. For our purposes, Eq. (3) is equivalent to Eq. (2).

In a diploid organism the transcriptional machinery of each gene X_i is composed of two alleles, each allele residing in one of the two chromosomes and transcribing mRNA at a certain rate which depends on the genotype and the concentrations of the gene's active transcription factors. Due to small differences in the two alleles' nucleotide sequences, transcription in the two alleles may proceed at (slightly) different transcription rates. The regulatory properties of the two products might also be different. However, a number of experimental results indicate that in many cases, the two alleles of a gene differ only in their regulatory domain without any variation in the coding region (Capon *et al.*, 2004; Chamary and Hurst, 2005; Duan and Antezana, 2003; Gehring *et al.*, 2001; Hoogendoorn *et al.*, 2003; Jones *et al.*, 2012; Mayo *et al.*, 2006; Peng *et al.*, 2005; Wang *et al.*, 1999; Rosenfeld *et al.*, 2005). In particular it seems reasonable to assume that for a homozygous gene, the two identical alleles are regulated in the same way and produce identical mRNAs.

Let the amounts or concentrations of mRNA produced by the two chromosomes of gene X_i be m_i^1 and m_i^2 , respectively. The total amount (concentration) of mRNA is $m_i = m_i^1 + m_i^2$. The same modelling approach as the one leading to Eq. (2) then gives

$$\begin{aligned}\dot{P}_i &= \rho_i m_i - \gamma_i P_i, \\ \dot{m}_i^1 &= R_i(P) - \mu_i m_i^1, \\ \dot{m}_i^2 &= R_i(P) - \mu_i m_i^2,\end{aligned}\tag{4}$$

where P is the vector of protein concentrations P_i . By addition this leads to

$$\begin{aligned}\dot{P}_i &= \rho_i m_i - \gamma_i P_i, \\ \dot{m}_i &= 2R_i(P) - \mu_i m_i.\end{aligned}\tag{5}$$

Applying the quasi-stationarity hypothesis to mRNA production leads finally to a single equation for the protein output concentration of node X_i :

$$\dot{z}_i = 2 \frac{\rho_i}{\mu_i} R_i(z) - \gamma_i z_i,\tag{6}$$

where again $z = P$. Eq. (6), or alternatively Eq. (3), is our final model for which we want to estimate the Jacobian J . The stationarity condition of Eq. (6),

$$2 \frac{\rho_i}{\mu_i} R_i(x) - \gamma_i x_i = 0,\tag{7}$$

will be denoted E_i as in (Plahte *et al.*, 2013).

Propagation functions, feedback functions and the Jacobian

For an investigation of how genetic variation at a locus propagates to the other loci in the network, it is easier and more fruitful to express all equilibrium values x_j as functions of the equilibrium value x_k of the perturbed node than to express them by the values of perhaps unknown parameters. We showed in Plahte *et al.* (2013) that the stationarity conditions of Eq. (1) define $n(n-1)$ *propagation functions* p_{jk} expressing how a change of the equilibrium value x_k for the locus X_k propagates via the network connections to any other locus X_j . The relation

$$x_j = p_{jk}(x_k, a^{(k)}), \quad j \neq k,\tag{8}$$

where $a^{(k)}$ is the set of parameters not specific to X_k , expresses an important property of p_{jk} ; for a given k , the propagation functions p_{jk} are invariant under genetic variation of X_k (Plahte *et al.*, 2013).

To determine the changes in any x_j , $j \neq k$, due to modification imposed on X_k , all we need is the resulting change in x_k and the propagation function p_{jk} . We do not need a model for how a genotypic change in X_k affects the rate function $r_k(z, a_k)$. This important property is a consequence of a theoretical result (Radulescu *et al.*, 2006) stating that the propagation function $p_{jk}(x_k, a^{(k)})$ is defined by all E_i except E_k , which is the only one containing the parameters a_k specific to X_k . It follows that the derivative q_{jk} of p_{jk} with respect to x_k can be easily estimated by the ratio of a small change in x_j divided by the change in x_k due to a small

genotypic variation in X_k . On the other hand, we showed in Plahte *et al.* (2013) that q_{jk} can be expressed in terms of the elements of the Jacobian J :

$$q_{jk}(x_k, a^{(k)}) = (-1)^{j+k} \frac{D^{(kj)}}{D^{(kk)}}, \quad j \neq k. \quad (9)$$

Accordingly, the propagation functions provide links between the observable protein concentrations and the Jacobian.

Because there are only $n(n-1)$ propagation functions, Eq. (9) alone is not sufficient to determine the n^2 elements of J completely from observable quantities. It does not tell how a genetic variation of X_k affects x_k itself. To determine this is a more difficult task because it requires knowledge of how a change of a parameter in a_k directly influences x_k , which in general may require a detailed model of the transcription and translation process. This information is in principle contained in the so-called *feedback functions* defined in Plahte *et al.* (2013).

Let $K = N \setminus \{k\}$. Expressing all x_j , $j \in K$, as $x_j = p_{jk}(x_k, a^{(k)})$ by means of all E_K and inserting this into E_k gives

$$\gamma_k x_k = r_k(x_k, \{p_{jk}(x_k, a^{(k)})\}_{j \neq k}, a_k). \quad (10)$$

The right-hand side of this equation is what we call the feedback function ϕ_k for X_k . The stationarity condition for X_k is then

$$\gamma_k x_k = \phi_k(x_k, a). \quad (11)$$

The feedback function ϕ_k describes and quantifies the feedback effects of changes in the equilibrium value of X_k on itself. If $\psi_k(x_k, a) = \phi'_k(x_k, a) \neq 0$, where the prime denotes the derivative with respect to x_k , there is an effective feedback of X_k on itself, mediated by one or more feedback loops. We showed in Plahte *et al.* (2013) that

$$\psi_k(x_k, a) = \frac{D}{D^{(kk)}} + \gamma_k. \quad (12)$$

Combining Eqs. (9) and (12) with the well-known formula for the matrix inverse in terms of determinant and minors, we get for $j \neq k$

$$(J^{-1})_{jk} = (-1)^{j+k} \frac{D^{(kj)}}{D} = q_{jk} \frac{1}{\psi_k - \gamma_k}, \quad (13)$$

or

$$JQ = C, \quad (14)$$

where Q is a square matrix defined by $Q_{ij} = q_{ij}$, and C is the diagonal matrix with diagonal elements $\psi_k - \gamma_k$, $k \in N$. It is easy to see that Eq. (13) is valid for $j = k$ as well because $q_{kk} = 1$.

Let us for a moment assume that Q is known and invertible while J and C are unknown. The effect of C is to multiply each row in Q^{-1} by a constant. Let $c = [c_1, \dots, c_n]$ be the nonzero diagonal elements of C . Because the system 1 given by $\dot{z}_i = c_i f_i(z)$, $i \in N$, has the same stationary states as system 2 given by $\dot{z}_i = f_i(z)$, while their Jacobians are related by $J_1 = CJ_2$, it might seem impossible to determine C from equilibrium values alone (see e.g. Sontag (2008)).

The problem is related to the fact that the invariance property of p_{jk} with respect to genotypic variation in X_k is not shared by the feedback function ϕ_k . The function itself changes, and unless this dependence is known, Eq. (11) cannot be used to estimate $\psi_k(x_k)$. Accordingly, it is not trivial to obtain an estimate of ψ_k by an arbitrary genotypic variation of X_k . If one were able to change some node-specific parameter in

X_k , the effect on x_k would depend explicitly on this parameter in a way that would require a model for the transcription and translation of the gene.

While our main object is to develop methods to estimate J , we note in passing that if J were known, Eq. (13) would yield an expression for Q . Because all $Q_{kk} = 1$, it follows that $C = (\text{diag}(J^{-1}))^{-1}$, hence

$$Q = J^{-1}(\text{diag}(J^{-1}))^{-1}. \quad (15)$$

This formula expresses the total derivative of any variable with respect to any other in terms of the partial derivatives of the dose-response functions and the degradation rates, taking all the network connections into account. This can be interpreted as follows: For the set of differentiable functions $f_i : R^n \rightarrow R^n$, $i = 1, \dots, n$, assume the set of equations $f_i(z) = 0$ has a solution x , and let its Jacobian be J , $J_{ij} = \partial f_i / \partial z_j|_{z=x}$. For each k , assume that the set of all the equations except $f_k(z) = 0$ define all x_j except x_k in terms of x_k in an open domain around x . (If all f_j have the particular form assumed in the present paper, this is ensured if a few additional conditions are fulfilled (Radulescu *et al.*, 2006).) Then for any j, k , $dx_j/dx_k = Q_{jk}$ is given by Eq. (15). The formula could be useful for computing the derivatives of functions defined implicitly by a set of equations.

Eq. (14) is the basis for reconstructing the Jacobian from observable equilibrium data. The main problem is to find a way to estimate C . Below we present two different approaches. One is to perturb the degradation rate γ_k which does not enter into ϕ_k , but enters into Eq. (11) in a known and simple way. The other approach is by allele knockout in diploid, homozygous genes, in which case we may assume that knocking out one of the alleles reduces the production rate of the gene by 50%.

Reconstruction of J by perturbing the mRNA degradation rates

We first analyse the problem of estimating J by perturbing the mRNA degradation rates μ_i in Eqs. (3) and (6) and recording the effects on the equilibrium concentrations. Note that μ_i occurs in the dose-response function of x_i in a known way. This will lead to an estimate \hat{J} of J .

Selecting one node X_k and keeping all other parameters fixed, we perturb the degradation rate μ_k from μ_k to $(1 + \omega)\mu_k$ and record the unperturbed equilibrium values x_j and the perturbed values $x_j^{[k]}$. The index in the bracket indicates which node has been perturbed. Because μ_k occurs in ϕ_k in a known way, we are able to derive a formula for ψ_k for any given k . Again we let $K = N \setminus \{k\}$. Expressing x_K in terms of their propagation functions p_{Kk} (which are defined by all E_j except E_k), the equation

$$\gamma_k x_k = \frac{\rho_k}{\mu_k} R_k(x_k, p_{Kk}(x_k)) = \phi_k(x_k) \quad (16)$$

defines x_k implicitly as a function of μ_k . Implicit differentiation with respect to μ_k gives readily

$$\frac{1}{\psi_k - \gamma_k} = \frac{\mu_k}{\gamma_k x_k} \frac{dx_k}{d\mu_k}.$$

Combined with Eq. (13) this yields

$$(J^{-1})_{jk} = \frac{\mu_k}{\gamma_k x_k} \frac{dx_k}{d\mu_k} q_{jk}. \quad (17)$$

Changing μ_k to $(1 + \omega)\mu_k$ induces a change from x_k to $x_k^{[k]}$. If $|\omega|$ and its effect on x_k are small, we can approximate the derivatives by

$$\frac{dx_k}{d\mu_k} \approx -\frac{\delta_k^{[k]}}{\omega\mu_k}, \quad q_{jk} \approx \frac{\delta_j^{[k]}}{\delta_k^{[k]}}, \quad (18)$$

where $\delta_j^{[k]} = x_j - x_j^{[k]}$ and $\delta_k^{[k]} = x_k - x_k^{[k]}$ are the effects on x_j and x_k of perturbing the degradation rate of X_k . This gives

$$(\hat{J}^{-1})_{jk} = -\frac{1}{\omega\gamma_k x_k} \delta_j^{[k]}, \quad (19)$$

$$\hat{J}H = -\omega B, \quad (20)$$

where \hat{J} is the estimated Jacobian, H is the square matrix with elements $H_{jk} = \delta_j^{[k]}$, and B is the diagonal matrix with diagonal elements $\gamma_k x_k$. If the μ_k are perturbed by different values ω_k , we get

$$\hat{J}H = -\Omega B, \quad (21)$$

where Ω is the diagonal matrix with $\Omega_{kk} = \omega_k$.

A convenient property of Eqs. (20) and (21) is that if the sign of each ω_k is known, the sign of the elements in \hat{J} can be determined even if the values of the γ_k are unknown, because all diagonal elements in B are positive and $\text{sign}(\Omega)$ would be known. In other words,

$$\text{sign}(\hat{J}) = -\text{sign}(\Omega)\text{sign}(H^{-1}). \quad (22)$$

The advantage of this approach from the mathematical point of view is that the degradation rates can be perturbed by different and small amounts. If there is no noise in the data, the induced errors when derivatives are approximated by ratios of finite differences can be made arbitrarily small. Nevertheless, errors in \hat{J} may occur if some nonzero elements in J are very small, or if H is very close to singular. In the latter case, arbitrarily large errors may occur in \hat{J} no matter how small ω is.

If recordings for several perturbed values of μ_k can be obtained, more precise estimates of the derivatives could be computed by more advanced mathematical methods. This is a great advantage of this method compared to the allele knockout method considered in the next subsection.

Reconstruction of J by allele knockouts in diploid loci

The allele knockout method is based on the reasonable assumption that if one of the alleles is knocked out, the production rate of the gene for fixed amounts of its regulators is reduced to one half its unperturbed value. Our starting point is again Eq. (13), where J now is the Jacobian of the system Eq. (6). However, in the derivation of (13) ϕ_k is derived from r_k , the dose-response function for a single allele. As the dose-response function for the homozygous diploid gene X_k is $2r_k(z)$, Eq. (13) must be replaced by

$$(J^{-1})_{jk} = (-1)^{j+k} \frac{D^{(kj)}}{D} = q_{jk} \frac{1}{2\psi_k - \gamma_k}. \quad (23)$$

To find an approximation to $\psi_k = d\phi_k/dx_k$ we use that the unperturbed and knock-out values x_k and $x_k^{[k]}$ are the solutions of

$$\begin{aligned}\gamma_k x_k &= 2\phi_k(x_k), \\ \gamma_k x_k^{[k]} &= \phi_k(x_k^{[k]}).\end{aligned}\tag{24}$$

Eqs. (24) give

$$\gamma_k \delta_k^{[k]} = 2\phi_k(x_k) - \phi_k(x_k^{[k]}) = 2\phi_k(x_k) - \phi_k(x_k - \delta_k^{[k]}).$$

If $\delta_k^{[k]}$ is small compared to x_k , expanding the last term to first order leads to

$$\frac{1}{2\psi_k - \gamma_k} = -\frac{1}{\gamma_k x_k^{[k]}} \delta_k^{[k]}.\tag{25}$$

Combined with Eq. (18) and Eq. (23) this finally gives

$$(\hat{J}^{-1})_{jk} = -\frac{1}{\gamma_k x_k^{[k]}} \delta_j^{[k]},\tag{26}$$

$$\hat{J}H = -\tilde{B},\tag{27}$$

where \tilde{B} is the diagonal matrix with diagonal elements $\gamma_k x_k^{[k]}$, and H was defined in the previous subsection. Note the similarity between Eqs. (20) and (27). Also note that even if the protein degradation rates are unknown, the dummy values $\gamma_k = 1$ will give the same sign to the elements of \hat{J} as Eq. (27).

Conditions on J

One should check that each eigenvalue of \hat{J} has a negative real part. When J is estimated by allele knockout, a further condition on the estimated Jacobian follows from Plahte *et al.* (2013). For a homozygous locus X_k the allele interaction value (Gjuvslund *et al.*, 2010) is defined by

$$\Delta_k = x_k - 2x_k^{[k]}.\tag{28}$$

Let F_k be the sum of all the terms in D in which there is a real regulation of X_k . This definition implies that F_k does not contain γ_k , rather, each term in F_k contains a factor representing a regulation of X_k , i.e. a factor $\partial r_k / \partial x_j$ for some j . Each term in F_k is a loop product of a feedback loop (called circuit in Plahte *et al.* (2013)) in J . Then

$$(-1)^n F_k \Delta_k < 0.\tag{29}$$

If P is any of these loop products, its contribution to F_k is $(-1)^v P$, where the signature factor v is the number of subloops in P with an even number of elements. If $(-1)^v P$ for some loop has the same sign as F_k , the loop is sign dominant, and

$$(-1)^{n+v-1} P \Delta_k > 0.\tag{30}$$

An estimate \hat{F}_k of F_k can be computed by analysing the loop structure of \hat{J} , and Δ_k can be computed directly from the observed equilibrium values. If the signs do not match with Eq. (29) or Eq. (30), there could be an error in the loop structure of \hat{J} , or the sign of F_k or Δ_k could be wrong due to noise. It is not obvious how to extract the most useful and reliable information from these inequalities. In the present paper we have made no effort to check our simulation results against these sign rules.

Discrepancy measures of estimated Jacobians

In tests of the method on systems with a known true Jacobian J , the estimate \hat{J} could be compared numerically to J in many ways to produce a measure of how well J has been reconstructed. Due to the degradation terms in the rate functions, the diagonal elements of J are generally nonzero except in the unlikely case that a positive autoregulation cancels the degradation term $-\gamma_j$. Because our method presupposes that the γ_j are known, we can in fact decide whether a nonzero diagonal element in \hat{J} indicates an autoregulation or only linear degradation. For this reason, it is more informative to compare $\hat{K} = \hat{J} + \Gamma$ with $K = J + \Gamma$, where Γ is the diagonal matrix with $\Gamma_{jj} = \gamma_j$.

The choice of error measure should reflect the main purpose of the reconstruction. In our view, the main objective is to reveal the connectivity of the gene network, while the dynamic properties are of less interest because the true system is most likely highly nonlinear. Our prime concern is therefore whether \hat{J} reproduces the right sign structure of J , using the sign function $\text{sign}(x) = x/|x|$ if $x \neq 0$, $\text{sign}(0) = 0$. Predicting the correct numeric magnitude of the matrix elements comes as a subordinate goal. We used the following classification for an estimated element \hat{K}_{ij} when the true value K_{ij} is known. If $\hat{K}_{ij} = 0$, it is called a *true zero* when $K_{ij} = 0$ and *false zero* when $K_{ij} \neq 0$. If $\hat{K}_{ij} \neq 0$, it is called a *true nonzero* when $K_{ij} \neq 0$ and $\text{sign}(\hat{K}_{ij}) = \text{sign}(K_{ij})$, a *false nonzero* when $K_{ij} = 0$ and a *wrong sign* when $K_{ij} \neq 0$ and $\text{sign}(\hat{K}_{ij}) \neq \text{sign}(K_{ij})$.

Our priorities are reflected in the discrepancy measure matrix M defined by the squared relative difference

$$M_{ij} = M_{ij}(\hat{K}, K) = \frac{(\hat{K}_{ij} - K_{ij})^2}{(|\hat{K}_{ij}| + |K_{ij}|)^2 + \varepsilon}, \quad i, j \in N. \quad (31)$$

The very small positive number ε , much less than the desired accuracy, is included to make the definition valid also if both elements are zero. It is a matter of elementary algebra to show that M_{ij} satisfies the requirements of a distance measure in R . Obviously, $M_{ij} = 0$ for a correct estimate, while $M_{ij} \approx 1$ for a false nonzero or a false zero in \hat{K}_{ij} or if \hat{K}_{ij} comes with the wrong sign. In all other cases, $0 < M_{ij} < 1$, a small value indicating a good estimate. If the error is small, M_{ij} is approximately equal to half the relative error in \hat{J}_{ij} . The average discrepancy measure is

$$\overline{M}(\hat{K}, K) = \frac{1}{n^2} \sum_{i,j \in N} M_{ij}(\hat{K}, K). \quad (32)$$

Simulation results

Estimating J by modifying mRNA degradation rates

When the equilibrium concentrations are noise-free and \hat{J} is estimated by perturbing the mRNA degradation rates, it should in theory be equal to J within computational accuracy, which depend on the user-defined relative perturbation ω of μ_k , integration tolerance, etc. Numerical simulations on randomly generated gene regulatory networks (see below) show that choosing ω of the order 10^{-2} to 10^{-3} gives a discrepancy measure of the same order of magnitude. By reducing ω sufficiently, \overline{M} can in theory be brought down to zero.

With real data, however, this is unattainable due to noise and experimental inaccuracy. Too small values of ω will lead to large uncertainties in the estimates of the derivatives. To obtain better estimates for the derivatives a fairly large number of observations with different perturbation levels would be needed. Many methods to estimate derivatives from noisy data can be found in the literature. Considering this to be part of the experimental and data processing setup, we do not elaborate this point any further. Instead, we limit ourselves to assuming that a single perturbation level is used, and that this experiment is repeated a certain number of times to produce a distribution of observed values. With this approach one must seek an optimal tradeoff between reducing the error in the derivatives due to finite differences (using a small ω) and minimising the noise to signal ratio. The differences in the equilibrium concentrations should not vary too much due to the noise, while ω should not be so large that nonlinear effects jeopardise the estimates of the derivatives. Intuitively, the perturbation level ω should be considerably larger than the standard deviation of the distributions of equilibrium levels.

Estimating Jacobian for the segment polarity network by means of allele knockouts

With its crude estimate of the derivatives it is less obvious that the allele knockout approach will work with an acceptable accuracy. To test this, we applied this method to the single cell segment polarity network (Tegnér *et al.*, 2003). We computed the protein equilibrium values by integrating the rate equations until the stable state x was reached with high accuracy. Employing total least squares, we used the equilibrium data to compute a matrix \hat{J}^0 from Eq. (27), then computed $\hat{K}^0 = \hat{J}^0 + \Gamma$, and finally subjected \hat{K}^0 to two kinds of cutoff to arrive at our final estimate \hat{K} (see Methods for details). To simulate the effect of noisy data, we also repeated the computation of \hat{K} after uniform noise had been added to x . More precisely, before estimating \hat{K} we added a noise term $Lu_i x_i$ to each equilibrium concentration x_i , where u_i was uniformly distributed in $[-1, 1]$ and L increased in steps of 0.05 from 0 to 0.25. For each noise level we ran $\ell = 50$ simulations.

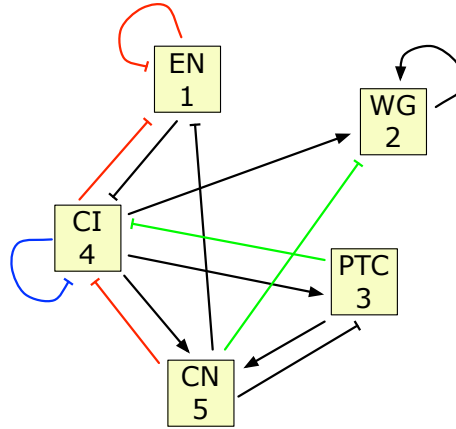


Figure 1: The protein connectivity network for the single cell segment polarity network according to Tegnér *et al.* (2003). Arrow heads signify activation, crossbars inhibition. Colour codes refer to our estimated \hat{K} without noise. Black arcs are correctly predicted in \hat{K} with the right sign, blue arcs are predicted with the wrong sign, green arcs signify false zeros (predicting no arc where there is one), and red arcs signify false nonzeros (predicting an arc where there is none). Thus, in the correct graph, black, blue and green arcs should be included, red arcs excluded.

The true and the predicted network connections obtained when no noise has been added to the protein equilibrium data, are shown in Figure 1. The cutoff value $c_J = 0.003$ appeared to give the lowest number of false elements. The elements are colour coded to show false nonzero elements (red), false zero elements (green) and nonzero but false sign (blue). With the same colour coding the discrepancy measure is

$$M = \begin{pmatrix} \textcolor{red}{1} & 0 & 0 & \textcolor{red}{1} & 0.032 \\ 0 & 0.015 & 0 & 0.009 & \textcolor{green}{1} \\ 0 & 0 & 0 & 0.002 & 0.001 \\ 0.206 & 0 & \textcolor{green}{1} & \textcolor{blue}{1} & \textcolor{red}{1} \\ 0 & 0 & 0.014 & 0.001 & 0 \end{pmatrix} \quad (33)$$

with average value $\bar{M} = 0.251$. For noisy data with noise level up to 0.25 the results are similar, with roughly the same average discrepancy measure (see Appendix C). Apart from the false nonzeros, false zeros and false signs for which $M_{jk} = 1$, the estimates are all of the right order of magnitude.

Estimating Jacobian for randomly generated gene regulatory networks by allele knockouts

The segment polarity network model has fixed network structure and parameter values. To complement this we performed large-scale simulations of gene regulatory network models with varying number of genes ($n = 5, 10, 20$). For each network size we ran 100 Monte Carlo simulations, sampling network connectivities and wild-type parameter values. The model structure and simulation setup is explained in detail in the Methods section. For each of the 100 randomly generated systems we simulated $\ell = 25$ sets of single knockout experiments, added noise ($L = 0, 0.05, 0.1, 0.15$) to the steady state expression levels and computed \hat{K} by means of ordinary least squares, trimming the estimates by the cutoffs $c_J = 0.005$ and $c_S = 0.5$ (see Methods).

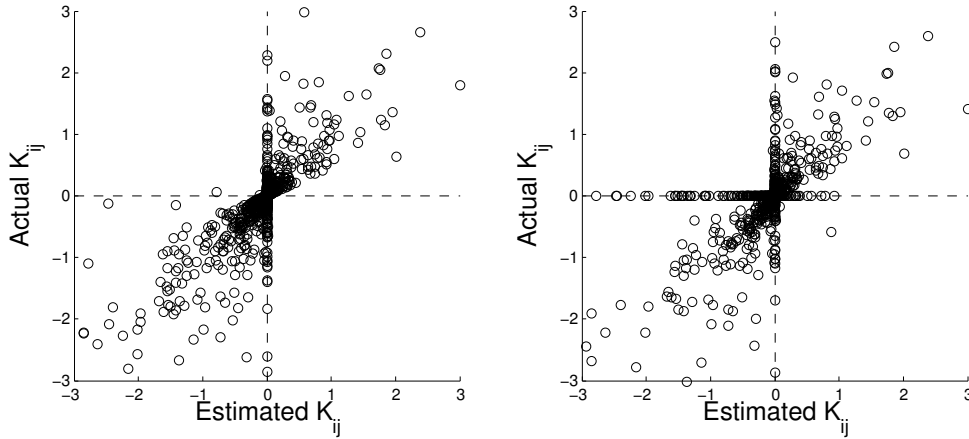


Figure 2: Scatterplots of \hat{K}_{ij} (x-axis) versus K_{ij} (y-axis) for *in silico* single-knockout experiments on 100 gene regulatory systems with $n = 10$ genes. The left panel shows results without noise on steady state expression levels while the right panel shows results from $\ell = 25$ repeated measurements with noise level $L = 0.1$. Observations where $\max\{|\hat{K}_{ij}|, |K_{ij}|\} > 3$ are not shown (68 and 85 of 10000 cases for the left and right panels, respectively).

Figure 2 shows scatterplots of estimated versus true values of elements in K for $n = 10$. For noise-free steady-state expression levels (left panel) we observe mainly *true nonzeros* in the 1st and 3rd quadrants and

false zeros on the K_{ij} -axis. When noise is added to the steady-state expression levels (right panel), false nonzeros appear on the \hat{K}_{ij} -axis. Estimates with wrong sign would appear in the 2nd and 4th quadrants, but are rarely seen. Similar patterns are observed for systems with $n = 5$ and $n = 20$ (see Appendix D).

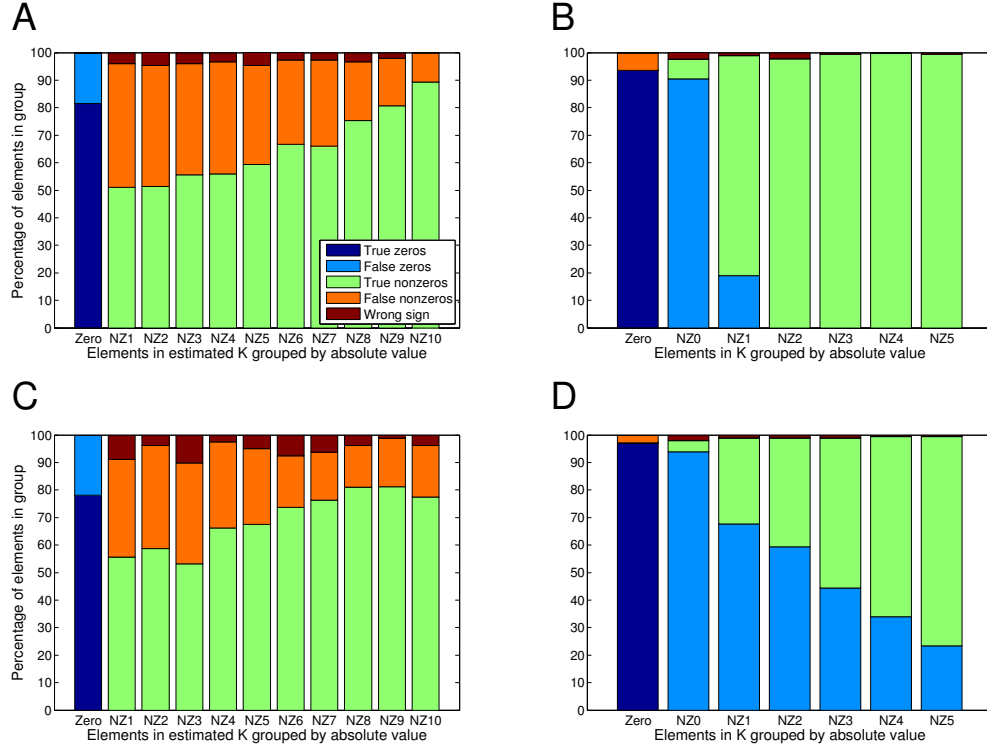


Figure 3: Summary of true and false discoveries of the signs of Jacobi elements for randomly generated gene regulatory networks with $n = 10$ genes. Each panel summarizes 10,000 (\hat{K}_{ij}, K_{ij}) pairs from *in silico* single-knockout experiments on 100 simulated systems. **(A)** Results for simulations without noise on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|\hat{K}_{ij}|$. The subset named Zero contains 8,495 pairs with $|\hat{K}_{ij}| = 0$, while the remaining pairs are sorted into 10 subsets NZ p , $p = 1, 2, \dots, 10$, with boundaries corresponding to the $(p - 1)$ th and p th 10-quantiles of the 1,505 $|\hat{K}_{ij}|$ values. **(B)** Results for simulations without noise on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|K_{ij}|$. The subset Zero contains 7,394 pairs with $|K_{ij}| = 0$, while the remaining pairs are sorted into 6 subsets; NZ0 which contains elements 1706 with $|K_{ij}| \leq c_J$ and NZ p , $p = 1, 2, \dots, 5$, with boundaries corresponding to the $(p - 1)$ th and p th 5-quantiles of the remaining 900 elements with $|K_{ij}| > c_J$. **(C)** Results for simulations with noise level $L = 0.1$ on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|\hat{K}_{ij}|$, as for **(A)**. The Zero subset contains 9,202 pairs. **(D)** Results for simulations with noise level $L = 0.1$ on steady state expression levels. The sorting of pairs is the same as in **(B)**.

Not visible in the origin of Figure 2 is a large number of *true zeros*, and the figure does not convey much information about the relative numbers of true and false sign estimates. Figure 3 gives an overview of this by subdividing the pairs of (\hat{K}_{ij}, K_{ij}) values into subsets based on their absolute values and displaying relative number of true and false positives for each subset. Using our methods on noise-free steady-state expression levels we find that slightly over 80% of the cases where $\hat{K}_{ij} = 0$ are *true zeros*, while the proportion of non-zero \hat{K}_{ij} that are *true non-zeros* increase with $|\hat{K}_{ij}|$ from 50% for the smallest estimates to 90% for the largest (Figure 3A). When dividing (\hat{K}_{ij}, K_{ij}) pairs into subsets based on K_{ij} (Figure 3B), we find that more

Table 1: Summary statistics of \overline{M} for simulated allele knockouts in gene regulatory network with varying number of genes (n) and noise level (L).

n	L	$\min(\overline{M})$	$\text{median}(\overline{M})$	$\text{mean}(\overline{M})$	$\max(\overline{M})$
5	0	0.177	0.384	0.393	0.671
5	0.1	0.233	0.421	0.423	0.77
10	0	0.165	0.236	0.242	0.394
10	0.1	0.173	0.241	0.244	0.363
20	0	0.09	0.133	0.137	0.229
20	0.1	0.102	0.137	0.139	0.182

than 90% of the zero elements in K are correctly identified as zeros. Furthermore, for the nonzero elements in K , false zeros are only a major problem in the group NZ0 where $|K_{ij}| \leq c_J$. The most important effects of adding noise to the steady-state expression values is (i) a considerable increase in the proportion of large elements in K that are incorrectly identified as zeros (Figure 3D, NZ1 to NZ5) and (ii) a reduction in the number of false non-zeros (Figure 3C, D). The overall pattern is similar for systems with $n = 5$ and $n = 20$ (see Appendix D).

The distribution of \overline{M} is shown in Table 1. The values of \overline{M} decrease with increasing number of nodes. The main reason is that the proportion of elements in K that are zero increases, and as seen in Figure 2 these zero elements are often correctly identified. For $n = 5$, equilibrium values with a low noise-level ($L = 0.1$) lead to slightly higher discrepancy measures, but for networks with 10 and 20 nodes there are no clear differences.

Conclusions

We have presented a method for estimating the Jacobian of an ODE model of a dynamic gene regulatory system based on the stable equilibrium values of the protein concentrations. The method is developed from our previous analysis of propagation of genetic variation (Plahte *et al.*, 2013). Together with known relative protein degradation rates, the observed shifts in the protein concentrations due to perturbations of the genes are sufficient to obtain an estimate \hat{J} of the true Jacobian J . We have analysed two experimentally feasible ways of perturbing the genes: perturbing the relative mRNA degradation rates in haploid or diploid systems, and allele knockout in diploid systems.

The reader should keep in mind that when one talks about a real system, the Jacobian always refers to a model of the system, not to the system itself. Our model framework is very general and includes explicit modelling of both mRNA and protein. We assume that the conversion rate from mRNA to protein is linear. However, as the Jacobian also is a linear representation valid around a stationary point of a usually nonlinear system, this linear conversion could be considered as a linear approximation of some nonlinear mRNA-protein response function.

Major advantages of our method are:

1. The estimate \hat{J} can be obtained from Eq. (20) or Eq. (27). Both equations are derived from Eq. (14). In this equation, Q may be estimated from the shifts in equilibrium values obtained by *any* kind of perturbation according to $q_{jk} \approx \delta_j^{[k]} / \delta_k^{[k]}$, c.f. Eq. (18). Knowing the relative degradation rates is not necessary. The effect of C is to multiply each row in H^{-1} by some factor proportional to the relative degradation rate γ_k of the corresponding protein. Thus, varying γ_k does not change the sign of the elements in \hat{J} , only their magnitudes. Accordingly, all the connectivities except autoregulations can be estimated with their right sign even if the γ_k are unknown. We consider this a major asset of the method.
2. If the protein relative degradation rates are known, C may be estimated by modifying the mRNA relative degradation rates or by allele knockouts. In those cases \hat{K} can be computed, a non-negative value of \hat{K}_{kk} pointing to an autoregulation in X_k .
3. Already existing knowledge about the action of one node on itself or other nodes can easily be incorporated. For example, if X_j is not autoregulated or negatively autoregulated, the value of C_{jj} should be such that $\hat{J}_{jj} < 0$. Only a sufficiently strong positive autoregulation may give $J_{jj} > 0$. Furthermore, if the sign of the action $X_k \rightarrow X_j$ is already known from experiments and the estimate of $(Q^{-1})_{jk}$ has the opposite sign, then necessarily $C_{jj} < 0$ if the sign of the estimate can be trusted. This fixes the sign of all the other nonzero elements in row j in \hat{J} . Such knowledge may also be used to set admissible ranges for the cutoffs.

Errors in \hat{J} could be due to observational noise or occur because derivatives have been approximated by ratios of finite differences. The effects of noisy data on the computation of \hat{J} have been analysed by Andreu *et al.* (2005). Because J is most likely sparse, in particular for large n , there will be a large number of false nonzero elements in \hat{J} before the cutoffs have been applied. If J is known, as it were in our numeric simulations, the values of the cutoffs c_j and c_s can be chosen to obtain an optimal fit. Our sign fluctuation analysis in terms of \bar{S} is of course quite simple-minded, and could be replaced or supplemented by more elaborate statistical analyses. In an experimental situation the optimal fit must instead be searched by repeated switching between experiment, theoretical analysis, and new, testable hypotheses based on high and low cutoff values.

A minimum value of c_j can always be set because too small values of \hat{J}_{jk} will be interpreted as indicating no effective regulatory action. However, the best cutoff value may be larger than this minimum. Large cutoffs gives fewer false nonzeros, but more false zeros (predicting a zero element in \hat{J} where the corresponding value in J is nonzero), and vice versa. Setting large cutoffs will lead to just a few nonzero elements in \hat{J} which most likely are correct, but a large number of false zeros. Similarly, setting small cutoffs will lead to near-certain zeros in \hat{J} , but many false nonzeros. The optimal cutoff values are a trade-off between these two extremes.

Even if the true Jacobian is unknown, the estimate \hat{J} could be subjected to a few tests. All the eigenvalues of the optimal \hat{J} should have negative real parts. This is not a strong criterion because the negative diagonal elements stemming from the degradation terms tend to make \hat{J} stable. A second consistency check is to use the sign conditions Eq. (29) relating the sign of the allele interaction values to the dominant loops.

Estimating derivatives from noisy data is notoriously difficult and error-prone. However, if the genes may

be perturbed by small amounts and with many different values (for instance by using RNA interference), more elaborate methods could be used to estimate the derivatives q_{jk} , even when data are noisy. The optimal method would depend on available experimental data, and is therefore outside the scope of this paper. Allele knockout, on the other hand, does not admit these options. An allele is either present or knocked out, leaving the researcher with just two data points from which the derivative may be approximated by finite differences. Accordingly, allele knockout is inherently a less precise method than degradation rate perturbation. For this reason, our simulations and discussion are mainly devoted to this suboptimal method in an attempt to determine its potential. Of course, combinations of the two methods could also be envisaged. Thus, our statistical analyses of the equilibrium data should be considered more as examples than recipes on how the data should be analysed. Our object has not been to develop the optimal way of analysing specific data, but to illustrate different approaches and to show that our method actually works.

Using steady state data, our method gives a sign estimate of the Jacobian, and if the protein degradation rates are known, a value estimate. Some connections are predicted more reliably than others, and should point to further experiments that may resolve the inconsistencies and uncertainties in the more uncertain estimates. However, with real data it is not to be expected that one reconstruction method alone will give the complete and correct answer. Where one method is inaccurate or fails, another may work, perhaps leading to conflicting conclusions that have to be resolved by further experimentation and analysis.

Networks and models

We tested both approaches for J reconstruction—allele knockouts and degradation rate manipulation—on *in silico* networks. In all the networks considered, protein and mRNA are modelled separately according to our general model framework. The stable protein concentration values were obtained by numeric integration of the rate equations until convergence to a steady state. In a few cases the solution approached a limit cycle. These cases were discarded. We compared \hat{J} estimated by the approaches described above to the “true” Jacobian J estimated directly from the rate equations in the standard numeric way by estimating the partial derivatives of the rate functions in the reduced model Eq. (3) (without the factor 2 if the system is not diploid).

Segment polarity network

One of the systems we used to test the method is the segment polarity network model of von Dassow *et al.* (2000) as it was adapted to a single cell (Tegnér *et al.*, 2003). In this model m_i and P_i are the mRNA and protein concentrations of the genes *engrailed* (*en*) ($i = 1$), *wingless* (*Wnt*) ($i = 2$), *Patched* (*Ptc*) ($i = 3$), *cubitus interruptus* (*ci*) ($i = 4$), and repressor fragment of *cubitus interruptus* ($i = 5$), respectively. Application of the quasi-stationarity hypothesis to the equations of motion given in Tegnér *et al.* (2003, see Supplement), leads to a realisation of Eqs. (3) for all $i \neq 4$ and

$$\dot{z}_4 = \frac{\lambda_4}{\mu_4} r_4(z) - r_5(z) - \gamma_4 z_4. \quad (34)$$

Here $z \in R^5$ is the vector of protein concentrations, r_i are the mRNA dose-response functions, and λ_i , γ_i , and μ_i are constant parameters (see Tegnér *et al.* (2003) for explicit formulae, parameter values and other

details). Due to the presence of r_5 in the equation for \dot{z}_4 , this system does not fit completely with our assumptions. A genotypic variation in X_5 will affect the dose-response function of z_4 directly, implying that the parameters describing the genotype of X_5 are not completely node-specific. It is interesting to see whether this fact will jeopardise our reconstruction or the system's Jacobian.

Random network models

As a further test we ran a series of numerical simulations with and without noise on a range of systems defined by Eqs. (6) of varying dimension and feedback structure for randomly generated dose-response functions and parameter values. For each system size ($n = 5, 10, 20$) we sampled 100 systems. For each system we first set up the overall connectivity as follow: for each node X_i we sampled two regulator nodes X_j, X_k , $i, j, k \in N$, and a mode of regulation (activation or repression) for each regulator. This simplifies the rate equation Eq. (6) to

$$\dot{z}_i = 2 \frac{\rho_i}{\mu_i} R_i(z_j, z_k, a_i) - \gamma_i z_i. \quad (35)$$

We used the dose-response function

$$R_i(z_j, z_k) = \alpha_i + \beta_i B_i(S_{ij}(z_j), S_{ik}(z_k)), \quad (36)$$

where α_i is the basal and $\alpha_i + \beta_i$ the maximal mRNA production rate, and B_i is the algebraic equivalent of a Boolean AND or OR function. We set $S_{ij}(z_j) = H(z_j, \theta_{ij}, p_{ij})$ if X_j activates X_i and $S_{ij}(z_j) = 1 - H(z_j, \theta_{ij}, p_{ij})$ if X_j represses X_i , where H is the Hill function $H(z, \theta, p) = z^p / (z^p + \theta^p)$ with threshold θ and steepness p . We sampled parameter values uniformly in the following ranges: $\beta_i, \mu_i \in (0, 10)$, $\alpha_i, \rho_i, \gamma_i, \theta_{ij}, \theta_{ik} \in (0, 1)$ and $p_{ij}, p_{ik} \in (1, 5)$.

Estimation methods

Least squares

The question of how to work out the best estimate \hat{J} from allele knockout or degradation rate perturbation data is not trivial. In all cases we get conditions of the kind $\hat{J}H = G$, where H and G are square and G is diagonal. If there is just a single set of measurements and H is invertible, our estimated Jacobian is uniquely given by $\hat{J} = GH^{-1}$. If data are noisy and we have a set of observations leading to H_i and G_i , $i = 1, \dots, \ell$, it is a matter of statistics to decide on an optimal estimation procedure. Obvious options are ordinary least squares (OLS) and total least squares (TLS) (Markovsky and Van Huffel, 2007). OLS assumes no measurement error in H , but works well as long as the measurement errors are small (Montgomery *et al.*, 2012). In its simple form, TLS assumes equal variances in H and G , which is at best only approximately fulfilled in our case. We used TLS for the segment polarity network and OLS for the randomly generated systems. (See Appendix B for details on OLS and TLS.) We do not claim that these are the optimal estimation procedures. Rather, they should be considered as examples used to show that the method actually works. More sophisticated estimates could certainly be found, but considering this as a problem belonging to the experimental and data processing setup, we do not elaborate this point any further.

Cutoffs

Since real networks, in particular large ones, seem to have sparse Jacobians, \hat{J} will in general probably contain many elements with small absolute values. The question is then whether these are just the consequence of approximation errors in the estimation procedure, or correspond to weak couplings in the network. Our algorithm does not make any qualitative distinction between an element in the Jacobian with a small absolute value and a zero element. This is reasonable in light of the probabilistic nature of transcription suggested by Bintu *et al.* (2005), who express the value of the dose-response function by the binding probabilities of transcription factors and polymerase. However, in common deterministic models an action of one gene on another either exists or does not exist. To relate our estimated \hat{J} to this kind of models we therefore have to apply some kind of cutoff to small elements in \hat{J}^0 , the estimate without cutoffs computed by one of the above estimation procedures..

In our simulations we applied two cutoffs to $\hat{K}^0 = \hat{J}^0 + \Gamma$. First we set all elements in \hat{K}^0 with an absolute value less than a cutoff c_J to zero. Secondly, for the cases of noisy data, we used the estimates $\hat{K}_l^0 = \hat{J}_l^0 + \Gamma$ obtained by means of Eq. (20) or Eq. (27) for all ℓ data set to define the average sign matrix $\bar{S} = (1/\ell) \sum_{l=1}^{\ell} \text{sign}(\hat{K}_l^0)$. The elements $(\hat{K}_l^0)_{ij}$ that are consistently equal to zero or have strongly varying sign for varying l , will come out as zero in \bar{S} or with small absolute values. If $|\bar{S}_{ij}|$ is smaller than a chosen sign variation cutoff c_S , the corresponding element in K is most likely zero. In these cases we set $\hat{K}_{ij} = 0$. In simulation studies when the true J is known, an optimal c_S can be found by comparing \hat{K} with K for different c_S . In real cases, a combination of intuition and repeated experiments may help in choosing the optimal value.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EP conceived the study and carried out the mathematical analysis. ABG designed and ran the numerical simulations and performed the statistical analysis. Both authors took part in evaluating the results and drafting the manuscript, and approved the final version.

Acknowledgements

We thank Stig W. Omholt for encouragement and comments. This work has been supported in part by The Research Council of Norway, project number 178901/V30, "Bridging the gap: disclosure, understanding and exploitation of the genotype-phenotype map".

Appendices

The appendices contain details of the singular perturbation analysis of the mRNA-protein system and of all derivations and details of the simulation procedure that are not included in the main document, and additional simulation results that are too lengthy to be included in the main document.

A mRNA and protein networks

The segment polarity network model of von Dassow *et al.* (2000) analysed in the paper is an example of a model framework in which the concentrations of mRNA and protein for each gene in the network are modelled independently. In this section we show how such models can be reduced by singular perturbation theory. For a network of n genes the model framework is Eq. (2), repeated here for convenience:

$$\begin{aligned}\dot{P}_i &= \rho_i m_i - \sigma_i P_i, \\ \dot{m}_i &= R_i(P) - \mu_i m_i.\end{aligned}$$

Here P_i and m_i are the concentrations of protein and mRNA of gene number i , respectively, R_i is the production rate (dose-response function) of mRNA, dependent on the concentration of the input proteins, ρ_i is the mRNA-protein conversion rate, and σ_i and μ_i are positive relative degradation rates. The gene products might act directly as transcription factors, or the function $R_i(P)$ might implicitly contain chains of reactions from the gene products to the real transcription factors so that R_i is the combined effect of these chains and the transcription. This framework has been used by a number of authors, see e.g. Polynikis *et al.* (2009) for a review.

As mRNA molecules are in general less stable than the corresponding protein molecules, we can safely assume that for all i , $\sigma_i \ll \mu_i$. We define $\varepsilon = \max\{\sigma_i/\mu_i\}$. Then $\varepsilon \ll 1$, and by a suitable renumbering of the genes we can always achieve $\varepsilon = \sigma_1/\mu_1$.

Using standard terminology in singular perturbation theory we call Eq. (2) *the full model*. To introduce ε in the equations we transform them to non-dimensional form by scaling the variables and the time t according to

$$\begin{aligned}P_i &= \frac{\rho_i}{\sigma_i \mu_i} x_i, \\ m_i &= \frac{1}{\mu_i} y_i, \\ t &= \frac{1}{\sigma_1} T.\end{aligned}\tag{A.1}$$

For convenience we continue to use a dot to denote time derivatives, but now with respect to the scaled time T . This leads to the dimensionless equations

$$\begin{aligned}\dot{x}_i &= \gamma_i y_i - \gamma_i x_i, \\ \varepsilon \dot{y}_i &= \eta_i R_i(x) - \eta_i y_i,\end{aligned}\tag{A.2}$$

where $\eta_i = \mu_i/\mu_1$, $\gamma_i = \sigma_i/\sigma_1$.

When ε is small and with a number of reasonable assumptions, we can make the quasi-stationarity hypothesis

$$y_i = R_i(x). \quad (\text{A.3})$$

This leads to *the reduced model*

$$\dot{x}_i = r_i(x) - \gamma_i x_i, \quad (\text{A.4})$$

where $r_i(x) = \gamma_i R_i(x)$. Obviously, the full and the reduced model have the same stationary states. The above derivation can be justified in a rigorous way by means of singular perturbation theory. In terms of the fast time $\tau = T/\varepsilon$, Eqs. (A.2) are

$$\begin{aligned} x'_i &= \varepsilon(\gamma_i y_i - \gamma_i x_i), \\ y'_i &= \eta_i R_i(x) - \eta_i y_i, \end{aligned} \quad (\text{A.5})$$

where the prime denotes differentiation with respect to τ . The crucial necessary assumption for the singular perturbation to be valid is that the stationary point of Eq. (A.5) is asymptotically stable for fixed values of x_i . In the present case this is ensured by assumption. Singular perturbation theory also ensures that when $\varepsilon \rightarrow 0$, the solution of the reduced system approaches the solution of the full system for all T except in a narrow, initial time interval.

B Estimate of J by least squares

In this section we consider briefly how to analyse the gene perturbation data using ordinary least squares (OLS) and total least squares (TLS). According to the main paper, both perturbation methods lead to

$$\hat{J}_l H_l = a \Gamma B_l = A_l, \quad (\text{B.1})$$

where the subscript $l = 1, \dots, \ell$ indicates the data obtained from experiment number l , and Γ is the diagonal matrix $\Gamma = \text{diag}(\gamma_k)$. For each l , $B_l = \text{diag}(x_k)$ in the case of degradation rate perturbation and $B_l = \text{diag}(x_k^{[k]})$ in the case of allele knockout, H is the matrix with elements $H_{jk} = x_j - x_j^{[k]}$, $a = -\omega$ in the case of degradation rate perturbation and $a = -1$ in the case of allele knockout. Finally, \hat{J}_l is the Jacobian estimated from dataset number l , and $A_l = a \Gamma B_l$.

The problem is to derive the best fit to J from the total dataset. Eq. (B.1) can be rewritten as

$$\Gamma(\hat{J}_l^{-1})^\top = \frac{1}{a} (H_l B_l^{-1})^\top. \quad (\text{B.2})$$

In this form the equation is amenable to an ordinary least squares solution because all noise is confined to the right-hand side. If Γ also has a nonzero variance, this is no longer so, and a more careful analysis is necessary. Assuming that Γ is known with negligible inaccuracy, we form the system $Y = G(\hat{J}^{-1})^\top$ by stacking all the right-hand sides of Eq. (B.2) into the $\ell n \times n$ matrix Y , and stacking ℓ Γ -matrices into the $\ell n \times n$ matrix G . Computing the least squares solution is straightforward and leads to

$$\hat{J}^{-1} = \frac{1}{\ell a} \sum_{l=1}^{\ell} \left(B_l^{-1} H_l^\top \right) \Gamma^{-1}. \quad (\text{B.3})$$

Combining this with Eq. (B.1) we readily arrive at

$$\hat{J}^{-1} = \frac{1}{\ell} \sum_{l=1}^{\ell} \hat{J}_l^{-1}, \quad (\text{B.4})$$

from which we find our final estimate by inversion.

The advantage of estimating \hat{J}^{-1} rather than \hat{J} is that the latter alternative would require the inverses of each H_l , while the above method essentially inverts the average of all the H_l , which has a lower variance.

To compute the TLS solution we define $H = [H_1, H_2, \dots, H_\ell]$ and $A = [A_1, A_2, \dots, A_\ell]$. Then Eq. (B.1) can be written $JH = A + E$ or

$$H^\top J^\top = A^\top + E^\top, \quad (\text{B.5})$$

where E is the error matrix. Because there are uncertainties in both H and A , TLS gives the optimal solution if all elements of H and A are normally distributed, uncorrelated and with equal variance. Assuming this is approximately correct, we find the TLS solution as follows (Markovsky and Van Huffel, 2007, Theorem 2). Let the dimensions of H and A be $n \times m$. We define C by $C = [H^\top, A^\top]$. Then the dimensions of C are $m \times 2n$. We may assume that $m \geq 2n$. Let a SVD decomposition of C be $C = USV^\top$, and let the singular values of C be $\sigma_1 \geq \dots \geq \sigma_{2n}$. We partition V as

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad (\text{B.6})$$

each partition having dimensions $n \times n$. Then there exists a TLS solution if and only if V_{22} is non-singular. In addition, the solution is unique if $\sigma_n \neq \sigma_{n+1}$. If both conditions are fulfilled, the TLS solution of Eq. (B.5) is $\hat{J}^\top = -V_{12}V_{22}^{-1}$, or

$$\hat{J} = (V_{22}^{-1})^\top V_{12}^\top. \quad (\text{B.7})$$

C Additional simulation results for the segment polarity network

The true and the predicted network connections obtained when no noise has been added to the protein equilibrium data, are shown in the main document. The true K is

$$K = \begin{pmatrix} 0 & 0 & 0 & 0 & -0.0562 \\ 0 & 0.0412 & 0 & 0.0090 & -0.0036 \\ 0 & 0 & 0 & 0.0168 & -0.0066 \\ -0.0523 & 0 & -0.9650 & -0.2565 & 0 \\ 0 & 0 & 0.9650 & 0.2565 & 0 \end{pmatrix}. \quad (\text{C.1})$$

For $L = 0$ (no noise) we found

$$\hat{K} = \begin{pmatrix} -0.0160 & 0 & 0 & -0.0153 & -0.0806 \\ 0 & 0.0527 & 0 & 0.0075 & 0 \\ 0 & 0 & 0 & 0.0184 & -0.0062 \\ -0.0196 & 0 & 0 & 0.0699 & -0.0327 \\ 0 & 0 & 1.2279 & 0.2381 & 0 \end{pmatrix}, \quad (\text{C.2})$$

using the cutoff value $c_J = 0.003$ which appears to be optimal. The elements are colour coded to show false nonzero elements (red), false zero elements (green) and nonzero but false sign (blue).

With the same colour coding the discrepancy measure is

$$M = \begin{pmatrix} \textcolor{red}{1} & 0 & 0 & \textcolor{red}{1} & 0.032 \\ 0 & 0.015 & 0 & 0.009 & \textcolor{green}{1} \\ 0 & 0 & 0 & 0.002 & 0.001 \\ 0.206 & 0 & \textcolor{green}{1} & \textcolor{blue}{1} & \textcolor{red}{1} \\ 0 & 0 & 0.014 & 0.001 & 0 \end{pmatrix} \quad (\text{C.3})$$

with average value $\bar{M} = 0.251$. Elements with false sign or false zeros/nonzeros in \hat{K} are equal to 1. For noisy data with noise level up to 0.25 the results are similar, with roughly the same average discrepancy measure, indeed somewhat smaller than without noise (Figure 4). Apart from the false nonzeros, false zeros and false signs for which $M_{jk} = 1$, the estimates are all of the right order of magnitude.

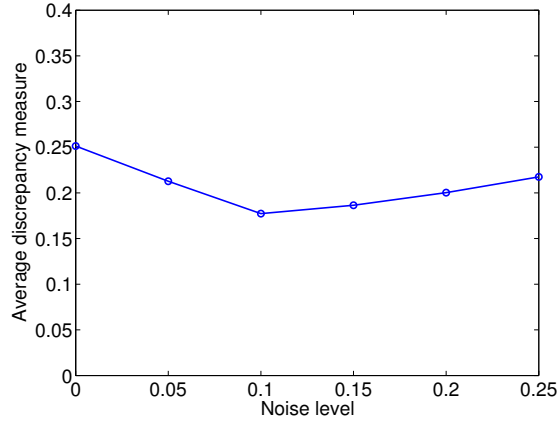


Figure 4: Average discrepancy measure \bar{M} of the segment polarity network for noise level L ranging from $L = 0$ to $L = 0.25$.

For $L = 0.10$ the average sign matrix \bar{S} is

$$\bar{S} = \begin{pmatrix} \textcolor{red}{-0.32} & \textcolor{red}{-0.20} & \textcolor{red}{0.08} & \textcolor{red}{-0.28} & -1 \\ \textcolor{red}{0.44} & 1 & \textcolor{red}{-0.08} & 1 & -0.92 \\ \textcolor{red}{0.20} & \textcolor{red}{-0.24} & \textcolor{red}{0.08} & 1 & -1 \\ -1 & \textcolor{red}{-0.20} & 0.04 & 1 & \textcolor{red}{-1} \\ \textcolor{red}{0.04} & \textcolor{red}{-0.20} & 1 & 1 & \textcolor{red}{0.20} \end{pmatrix}. \quad (\text{C.4})$$

The elements that are zero in K are shown in red. With two exceptions, these are the elements whose values in \bar{S} are close to zero. With the same exceptions, the remaining elements in \bar{S} have values close to ± 1 . The same is true for the other noise levels investigated. For all noise levels, the magnitude of the elements in \bar{S} are quite clearly separated in two classes, the smaller elements corresponding roughly to the fourteen zero elements in K (Figure 6).

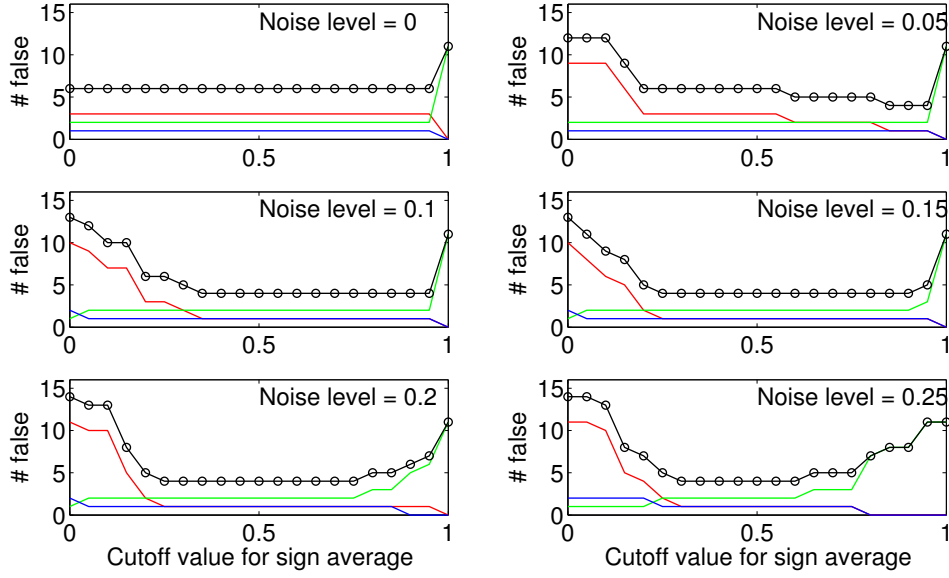


Figure 5: The number of false nonzero elements (red), false zero elements (green), false signs (blue) and the sum of all three (black with circles) in the estimated Jacobian for the segment polarity network for varying sign variation cutoff c_S , and noise level L ranging in steps of 0.05 from $L = 0$ (no noise) to $L = 0.25$.

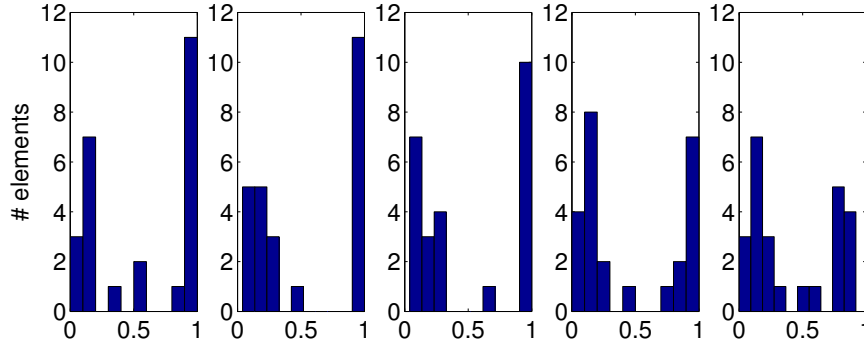


Figure 6: The distributions of the absolute values of the elements in \bar{S} for noise levels $L = 0.05$ (far left) in steps of 0.05 to $L = 0.25$ (far right).

D Additional simulation results for the randomly generated systems

In the main file we presented the results of the simulations on randomly generated systems with $n = 10$ genes. Here we present the corresponding results for $n = 5$ and $n = 20$. For interpretation of the diagrams see the legend to Figure 3.

When noise is added, the numbers of false elements also depend on the sign cutoff c_S . However, for most noise levels, the value of c_S is not very critical. Figure 5 show the number of false nonzero elements, false zero elements, false signs and the sum of all three for varying c_S and a range of noise levels. For most noise levels the total number of false elements is four, corresponding to 84% correctly predicted elements, for a

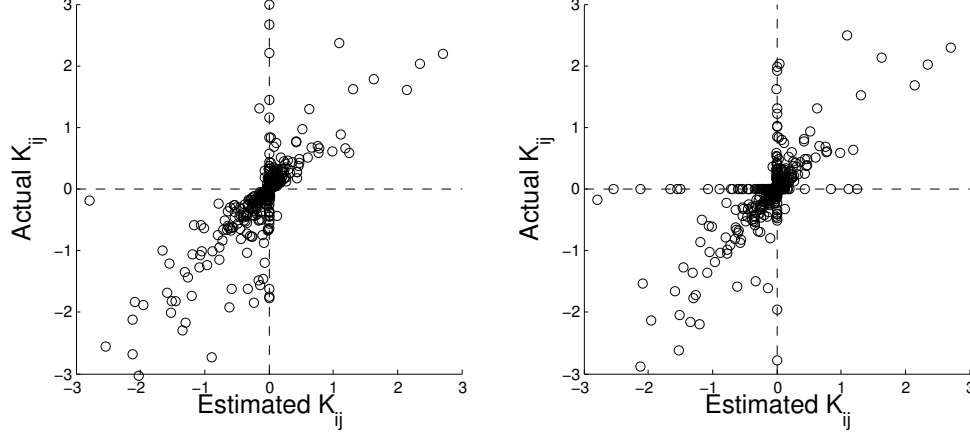


Figure 7: Scatterplots of K_{ij} (y-axis) versus \hat{K}_{ij} (x-axis) for *in silico* single-knockout experiments on 100 randomly generated gene regulatory systems with $n = 5$ genes. The left panel shows results without noise on steady state expression levels. The right panel shows results from $\ell = 25$ repeated measurements with noise level $L = 0.1$. Observations where $\max(|\hat{K}_{ij}|, |K_{ij}|) > 3$ are not shown (24 and 26 of 2,500 (\hat{K}_{ij}, K_{ij}) pairs for the left and right panel, respectively).

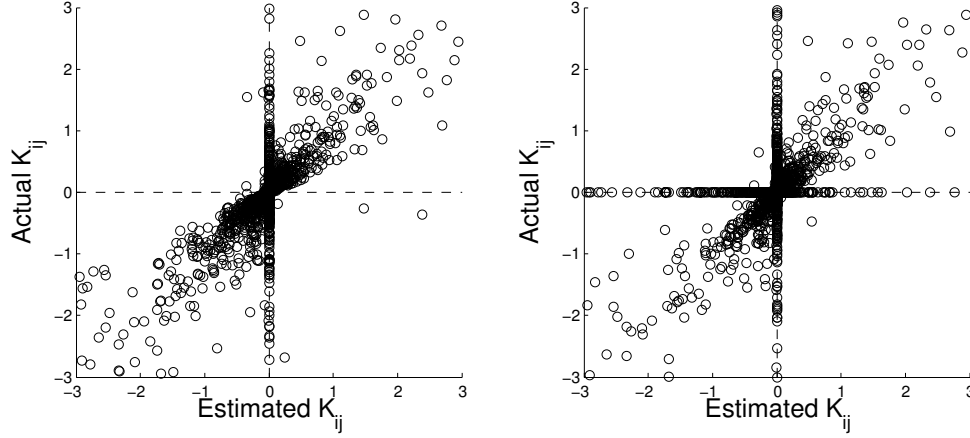


Figure 8: Scatterplots of K_{ij} (y-axis) versus \hat{K}_{ij} (x-axis) for *in silico* single-knockout experiments on 100 randomly generated gene regulatory systems with $n = 20$ genes. The left panel shows results without noise on steady state expression levels. The right panel shows results from $\ell = 25$ repeated measurements with noise level $L = 0.1$. Observations where $\max(|\hat{K}_{ij}|, |K_{ij}|) > 3$ are not shown (157 and 174 of 40,000 (\hat{K}_{ij}, K_{ij}) pairs for the left and right panel, respectively).

wide range of cutoff values.

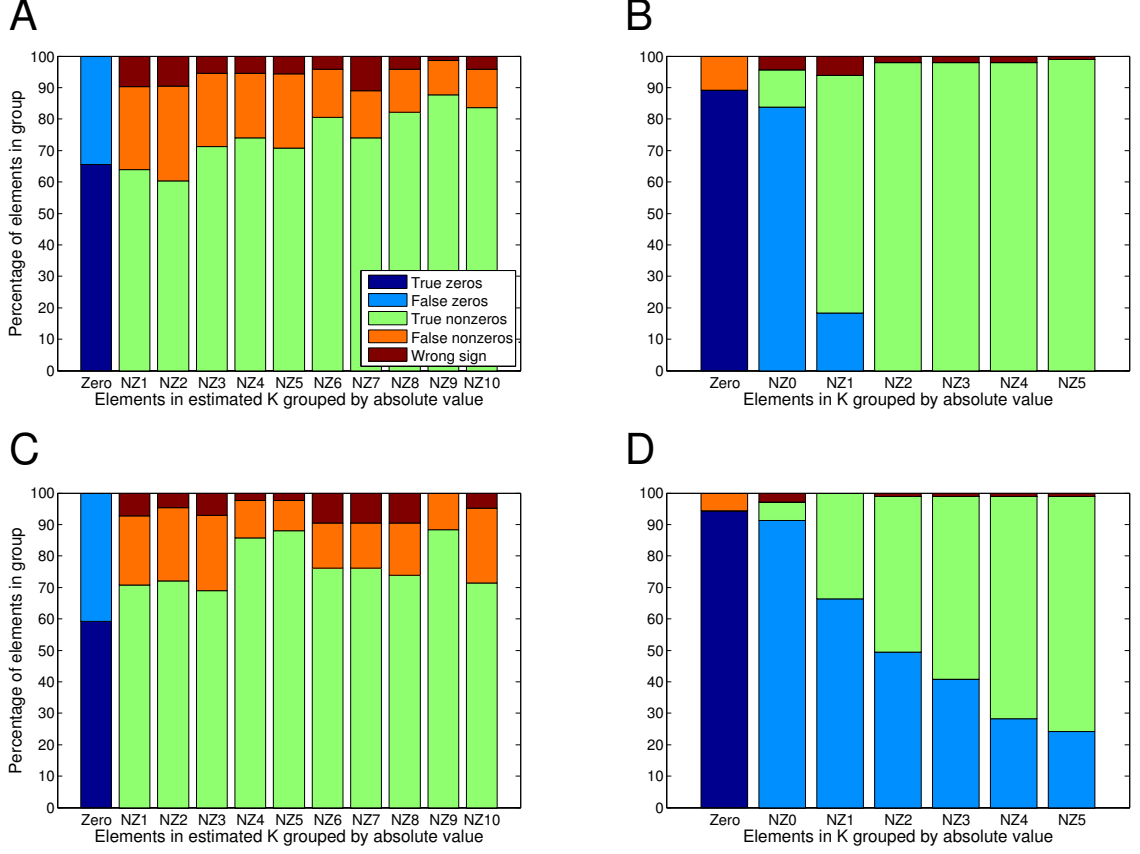


Figure 9: Summary of true and false discoveries of the signs of Jacobi elements for randomly generated gene regulatory networks with $n = 5$. Each panel summarizes 2,500 (\hat{K}_{ij}, K_{ij}) pairs from *in silico* single-knockout experiments on 100 simulated gene regulatory systems. **(A)** Results for simulations without noise on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|\hat{K}_{ij}|$. The subset named Zero contains 1,771 pairs with $|\hat{K}_{ij}| = 0$, while the remaining pairs are sorted into 10 subsets NZp , $p = 1, 2, \dots, 10$, with boundaries corresponding to the $(p - 1)$ th and p th 10-quantiles of the 729 $|\hat{K}_{ij}|$ values. **(B)** Results for simulations without noise on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|K_{ij}|$. The subset named Zero contains 2,077 pairs with $|K_{ij}| = 0$, while the remaining pairs are sorted into 10 subsets NZp , $p = 1, 2, \dots, 10$, with boundaries corresponding to the $(p - 1)$ th and p th 10-quantiles of the 423 $|K_{ij}|$ values. **(C)** Results for simulations with noise level $L = 0.1$ on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|\hat{K}_{ij}|$, as for **(A)**. The Zero group contains 2,006 pairs. **(D)** Results for simulations with noise level $L = 0.1$ on steady state expression levels. The sorting of pairs is the same as in **(B)**.

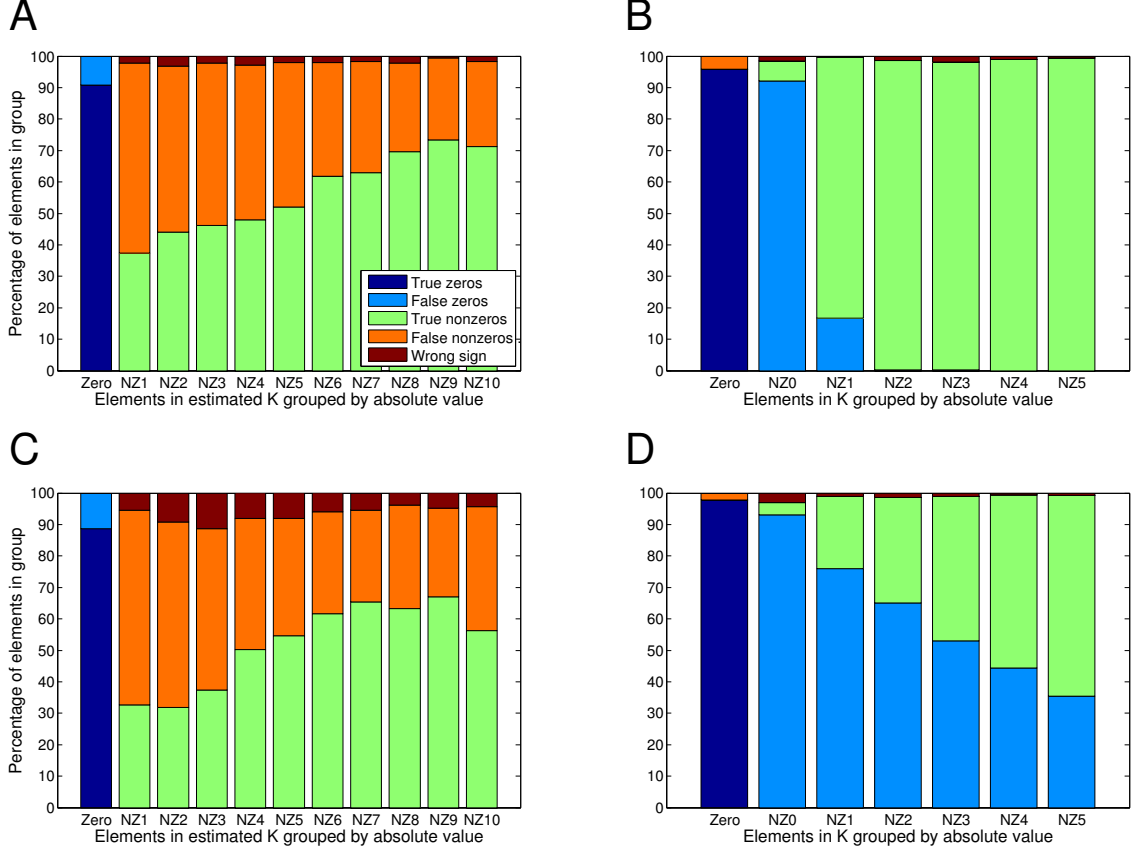


Figure 10: Summary of true and false discoveries of the signs of Jacobi elements for randomly generated gene regulatory networks with $n = 20$ genes. Each panel summarizes 40,000 (\hat{K}_{ij}, K_{ij}) pairs from *in silico* single-knockout experiments on 100 simulated systems. **(A)** Results for simulations without noise on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|\hat{K}_{ij}|$. The subset named Zero contains 36,462 pairs with $|\hat{K}_{ij}| = 0$, while the remaining pairs are sorted into 10 subsets NZ_p , $p = 1, 2, \dots, 10$, with boundaries corresponding to the $(p - 1)$ th and p th 10-quantiles of the 3,538 $|\hat{K}_{ij}|$ values. **(B)** Results for simulations without noise on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|K_{ij}|$. The subset Zero contains 38,149 pairs with $|K_{ij}| = 0$, while the remaining pairs are sorted into 10 subsets NZ_p , $p = 1, 2, \dots, 10$, with boundaries corresponding to the $(p - 1)$ th and p th 10-quantiles of the 1,851 $|K_{ij}|$ values. **(C)** Results for simulations with noise on steady state expression levels. The (\hat{K}_{ij}, K_{ij}) pairs are sorted into subsets (x-axis) based on $|\hat{K}_{ij}|$, as for **(A)**. The Zero subset contains 38,139 pairs. **(D)** Results for simulations with noise level $L = 0.1$ on steady state expression levels. The sorting of pairs is the same as in **(B)**.

References

- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat Rev Genet*, **8**(6), 450–461.
- Andrec, M., Kholodenko, B. N., Levy, R. M., and Sontag, E. (2005). Inference of signaling and gene regulatory networks by steady-state perturbation experiments: structure and accuracy. *Journal of Theoretical Biology*, **232**(3), 427–441.
- Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, **15**(2), 116–124.
- Brazhnik, P. (2005). Inferring gene networks from steady-state response to single-gene perturbations. *Journal of Theoretical Biology*, **237**(4), 427–440.
- Brazhnik, P., de la Fuente, A., and Mendes, P. (2002). Gene networks: how to put the function in genomics. *Trends in Biotechnology*, **20**(11), 467–472.
- Camacho, D., Licona, P. V., Mendes, P., and Laubenbacher, R. (2007). Comparison of reverse-engineering methods using an *in Silico* network. *Annals of the New York Academy of Sciences*, **1115**, 73–89.
- Capon, F., Allen, M. H., Ameen, M., Burden, A. D., Tillman, D., Barker, J. N., and Trembath, R. C. (2004). A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Human Molecular Genetics*, **13**, 2361–2368.
- Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., and Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, **48**, 55–65.
- Chamary, J. V. and Hurst, L. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology*, **6**, R75.
- Cho, K.-H., Choo, S.-M., Wellstead, P., and Wolkenhauer, O. (2005). A unified framework for unraveling the functional interaction structure of a biomolecular network based on stimulus-response experimental data. *FEBS Letters*, **579**(20), 4520–4528.
- Cho, K. H., Choo, S. M., Jung, S. H., Kim, J. R., Choi, H. S., and Kim, J. (2007). Reverse engineering of gene regulatory networks. *IET Systems Biology*, **1**(3), 149–163.
- de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, **9**(1), 67–104.
- Duan, J. and Antezana, M. A. (2003). Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. *Journal of Molecular Evolution*, **57**, 694–701.
- Emmert-Streib, F. and Dehmer, M. (2011). Networks for systems biology: conceptual connection of data and function. *IET Systems Biology*, **5**(3), 185–207.
- Emmert-Streib, F., Glazko, G., Gökmen, A., and De Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in Genetics*, **3**(8).

- Gehring, N. H., Frede, U., Neu-Yilik, G., Hundsdoerfer, P., Vetter, B., Hentze, M. W., and Kulozik, A. E. (2001). Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nature Genetics*, **28**(4), 389–392.
- Gjuvsland, A. B., Plahte, E., Ådnøy, T., and Omholt, S. W. (2010). Allele interaction – single locus genetics meets regulatory biology. *PLoS ONE*, **5**(2), e9379.
- Goutsias, J. and Lee, N. H. (2007). Computational and experimental approaches for modeling gene regulatory networks. *Current Pharmaceutical Design*, **13**(14), 1415–1436.
- Hoogendoorn, B., Coleman, S. L., Guy, C. A., Smith, K., Bowen, T., Buckland, P. R., and O'Donovan, M. C. (2003). Functional analysis of human promoter polymorphisms. *Human Molecular Genetics*, **12**, 2249–2254.
- Ichinose, N., Yada, T., Gotoh, O., and Aihara, K. (2008). Reconstruction of transcription-translation dynamics with a model of gene networks. *Journal of Theoretical Biology*, **255**(4), 378–386.
- Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., Brady, S. D., Zhang, H., Pollen, A. A., Howes, T., Amemiya, C., Lander, E. S., Di Palma, F., Lindblad-Toh, K., and Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**(7392), 55–61.
- Kholodenko, B. N., Kiyatkin, A., Bruggeman, F. J., Sontag, E., Westerhoff, H. V., and Hoek, J. B. (2002). Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(20), 12841–12846.
- Lewis, J. (2003). Autoinhibition with transcriptional delay: A simple mechanism for the zebrafish somitogenesis oscillator. *Current Biology*, **13**(16), 1398–1408.
- Markovsky, I. and Van Huffel, S. (2007). Overview of total least-squares methods. *Signal Processing*, **87**(10), 2283–2302.
- Mayo, A. E., Setty, Y., Shavit, S., Zaslaver, A., and Alon, U. (2006). Plasticity of the *cis*-regulatory input function of a gene. *PLoS Biology*, **4**, e45.
- Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley, New York.
- Peng, J., Murray, E. L., and Schoenberg, D. R. (2005). The poly(A)-limiting element enhances mRNA accumulation by increasing the efficiency of pre-mRNA 3' processing. *RNA*, **11**, 958–965.
- Plahte, E., Gjuvsland, A. B., and Omholt, S. W. (2013). Propagation of genetic variation in gene regulatory networks. *Physica D: Nonlinear Phenomena*, **256-257**, 7–20.
- Polynikis, A., Hogan, S. J., and di Bernardo, M. (2009). Comparing different ODE modelling approaches for gene regulatory networks. *Journal of Theoretical Biology*, **261**(4), 511–530.
- Radulescu, O., Lagarrigue, S., Siegel, A., Veber, P., and Le Borgne, M. (2006). Topology and static response of interaction networks in molecular biology. *J R Soc Interface*, **3**(6), 185–196.
- Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science*, **307**(5717), 1962–1965.

- Ross, J. (2008). Determination of complex reaction mechanisms. Analysis of chemical, biological and genetic networks. *Journal of Physical Chemistry A*, **112**(11), 2134–2143.
- Sontag, E., Kiyatkin, A., and Kholodenko, B. N. (2004). Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics*, **20**(12), 1877–1886.
- Sontag, E. D. (2008). Network reconstruction based on steady-state data. *Essays in Biochemistry: Systems Biology, Vol 45*, **45**, 161–176.
- Stark, J., Callard, R., and Hubank, M. (2003a). From the top down: towards a predictive biology of signalling networks. *Trends in Biotechnology*, **21**(7), 290–293.
- Stark, J., Brewer, D., Barenco, M., Tomescu, D., Callard, R., and Hubank, M. (2003b). Reconstructing gene networks: what are the limits? *Biochemical Society Transactions*, **31**, 1519–1525.
- Tegnér, J., Yeung, M. K. S., Hasty, J., and Collins, J. J. (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(10), 5944–5949.
- Tirosh, I. and Barkai, N. (2011). Inferring regulatory mechanisms from patterns of evolutionary divergence. *Mol Syst Biol*, **7**.
- von Dassow, G., Meir, E., Munro, E., and Odell, G. (2000). The segment polarity network is a robust development module. *Nature*, **406**(6792), 188–192.
- Wang, R. L., Stec, A., Hey, J., Lukens, L., and Doebley, J. (1999). The limits of selection during maize domestication. *Nature*, **398**, 236–239.
- Yalamanchili, N., Zak, D. E., Ogunnaike, B. A., Schwaber, J. S., Kriete, A., and Kholodenko, B. N. (2006). Quantifying gene network connectivity *in silico*: scalability and accuracy of a modular approach. *IEEE Proceedings Systems Biology*, **153**(4), 236–246.
- Yip, K. Y., Alexander, R. P., Yan, K.-K., and Gerstein, M. (2010). Improved reconstruction of *in silico* gene regulatory networks by integrating knockout and perturbation data. *PLoS ONE*, **5**(1), e8121.